eJHI
electronic Journal of Health Informatics
http://www.ejhi.net
2008; Vol 3(1): e6

Current Developments of *k*-Anonymous Data Releasing

Jiuyong Li[1], Hua Wang[2], Huidong Jin[3], Jianming Yong[4]
[1]School of Computer and Information Science,
University of South Australia, Mawson Lakes, Adelaide, Australia, 5095
[2]Department of Mathematics and Computing,
University of Southern Queensland,Toowoomba, Australia 4350
[3]NICTA Canberra Lab, 7 London Circuit, Canberra ACT 2601, Australia,
[4]School of Information Systems, University of Southern Queensland,
Toowoomba, Australia, 4350

Abstract

Disclosure-control is a traditional statistical methodology for protecting privacy when data is released for analysis. Disclosure-control methods have enjoyed a revival in the data mining community, especially after the introduction of the k-anonymity model by Samarati and Sweeney. Algorithmic advances on k-anonymisation provide simple and effective approaches to protect private information of individuals via only releasing k-anonymous views of a data set. Thus, the k-anonymity model has gained increasing popularity. Recent research identifies some drawbacks of the k-anonymity model and presents enhanced k-anonymity models. This paper reviews problems of the k-anonymity model and its enhanced variants, and different methods for implementing k-anonymity. It compares the k-anonymity model with the secure multiparty computation-based privacy-preserving techniques in the data mining literature. The paper also discusses further development directions of the k-anonymous data releasing.

Keywords: Privacy preserving, data releasing, k-anonymity

0. Feedbacks from ehPASS’06

This paper has attracted the interest of health researchers as a comprehensive review of new techniques for privacy preserving data publishing. Most discussed methods not only protect personal identifications in a published data set, but remove possible inference channels for private information. They make strong protections of privacy in data. However, a major concern is that the utility of published data is reduced if a strong model is employed. In most applications where data exchange is largely between different medical professionals, the k-anonymity model is sufficient. However, when data are published in a wide community, where malicious users may try to hack private information, the strong models are definitely useful. In addition, currently there is no legislation clarifying how strong protection is required. The application of new techniques also relies on the future legislative development.

1. Introduction

Various organisations, such as hospitals, medical administrations and insurance companies, have collected a large amount of data over years. However, gold nuggets in these data are unlikely to be discovered if the data is locked in data custodians’ storage. A major risk of releasing data for public research is revealing the private information of individuals in data.

Disclosure-control [1, 2, 3, 4] is a traditional approach for privacy-preserving data releasing. Most of them concentrate on maintaining statistical properties of data. Disclosure-control study attracts increasing interest the data mining community due to privacy concerns in powerful data mining processes. The data mining community aims to build strong privacy-preserving models and to design efficient, optimal and scalable heuristic solutions. Data perturbation [5, 6,

[1] This work has been done when author was with Department of Mathematics and Computing, University of Southern Queensland, Toowoomba.

The *electronic Journal of Health Informatics* is an international journal committed to scholarly excellence and dedicated to the advancement of Health Informatics and information technology in healthcare. ISSN: 1446-4381

7] and the *k*-anonymity model [8, 9, 10] are two major techniques to achieve the goal. Data perturbation methods are not for all but only for some specific data mining functionalities [5, 6, 7]. The *k*-anonymity model has been extensively studied recently because of its simplicity and general effectiveness, [11, 12, 13, 14, 15, 16].

An alternative methodology for privacy preservation is based on Secure Multiparty Computation (SMC) techniques [17, 18], where SMC is used to ensure that nothing should be revealed during the data mining process [19, 20, 21, 22]. In comparison with the disclosure-control data releasing methodology, the SMC-based methodology is usually inefficient. Disclosure control is more efficient and gives users more flexibilities for using data. Therefore, disclosure-control data releasing, such as *k*-anonymity, has a great potential for immediate real world applications.

In the rest of this paper, we introduce the *k*-anonymity model and discuss how it protects private information in data in Section 2. We then discuss problems associating with the *k*-anonymity model and some enhanced models to overcome these problems in Section 3. After that we summarise major techniques for implementing *k*-anonymisation in Section 4. In Section 5, we compare relative strengths and weaknesses of *k*-anonymisation with SMC-based privacy-preserving techniques. Finally, we conclude the paper and discuss some possible future directions.

2. *k*-anonymity model

Many organisations are increasingly sharing data by exchanging or publishing raw data containing unaggregated information about individuals. The data is normally de-identified. Names, medical care card numbers, and addresses are removed. It is assumed that individual is not identifiable and hence their privacy, such as medical conditions, is protected.

However, such a de-identification procedure does not guarantee the privacy of individuals in the data. Sweeney reported that 87% of the population of the United States can be uniquely identified by the combinations of attributes: gender, date of birth, and 5-digit zip code [23]. Sweeney also showed that the medical records of the governor of Massachusetts are supposedly anonymous but his medical data are uniquely identified by a linking attack. Gender, date of birth and zip code attributes were used in the linking attack by linking Massachusetts voter registration records, which include name, gender, zip code, and date of birth, to medical records, which include gender, zip code, date of birth as well as medical conditions.

Though explicit identifiers are removed from a data set, some attributes, for example, gender, date of birth and postcode in the above example, can potentially identify individuals in populations. Such a set of attributes is called a *quasi-identifier*. Samarati and Sweeney proposed a model for privacy protection called *k-anonymity* [8, 23, 10]. A data set satisfies *k*-anonymity if every record in the data set is identical to at least (*k* – 1) other records with respect to the set of quasi-identifier attributes; and such a data set is so-called *k-anonymous*. As a result, an individual is indistinguishable from at least (*k* – 1) individuals in a *k*-anonymous data set.

	Quasi identifier			Other attributes	Sensitive attributes
	Gender	Age	Postcode		Diagnosis
1	male	25	4350	… … …	depression
2	male	27	4351	… … …	depression
3	male	22	4352	… … …	flu
4	male	28	4353	… … …	flu
5	female	34	4352	… … …	depression
6	female	31	4352	… … …	Flu
7	female	38	4350	… … …	alcohol addiction
8	female	35	4350	… … …	alcohol addiction
9	m ale	42	4351	… … …	alcohol addiction
10	male	42	4350	… … …	alcohol addiction
11	male	45	4351	… … …	alcohol addiction
12	male	45	4350	… … …	alcohol addiction

Table 1. A raw medical data set

For example, Table 1 shows a simplified medical data set. It does not contain personal identification attributes, such as name, address, and medical care card number. However, the unique combinations of gender, age and postcode still reveal sensitive information of individuals. For instance, the first record is unique in these three attributes, and the patient is potentially identifiable. As a result, depression condition of the patient may be revealed by unique combinations, such as records 1, 2 and 5.

To avoid privacy breaching, Table 1 can be modified to Table 2. In Table 2, age is grouped into intervals, and postcodes are clustered into large areas. Symbol '*' denotes any digit. A record in the quasi-identifier is identical to at least 3 other records, and therefore, no individual is identifiable.

k-anonymisation becomes popular in data publishing because of its simplicity and the availability of many algorithms. However, the *k*-anonymity model may still reveal sensitive information under some attacks, and hence does not guarantee privacy. We discuss its enhanced models in the following section.

3. Enhanced *k*-anonymity models

The *k*-anonymity model may reveal sensitive information under the following two types of attacks [24].

1. Homogeneity attack to a *k*-anonymity table: Bob and Tom are two hostile neighbours. Bob knows that Tom goes to hospital recently and tries to find out the disease Tom suffers. Bob finds the 4-anonymous table as in Table 2. He knows that Tom is 42 years' old and lives in the suburb with postcode 4350. Tom must be record 9, 10, 11, or 12. All four patients are alcohol addiction sufferers. Bob knows for sure that Tom suffers alcohol addiction.

Therefore, homogeneous values in the sensitive attribute of a *k*-anonymous group leak private information.

	Quasi identifier			Other attributes	Sensitive attributes
	Gender	Age	Postcode		Diagnosis
1	male	20-29	435*	… … …	depression
2	male	20-29	435*	… … …	depression
3	male	20-29	435*	… … …	flu
4	male	20-29	435*	… … …	flu
5	female	30-39	435*	… … …	depression
6	female	30-39	435*	… … …	flu
7	female	30-39	435*	… … …	alcohol addiction
8	female	30-39	435*	… … …	alcohol addiction
9	male	40-49	435*	… … …	alcohol addiction
10	male	40-49	435*	… … …	alcohol addiction
11	male	40-49	435*	… … …	alcohol addiction
12	male	40-49	435*	… … …	alcohol addiction

Table 2. A 4-anonymous view of Table 1

2. Background knowledge attack to a *k*-anonymity table: Bob and Alice are friends and Bob does not want Alice to know his medical condition. Alice knows Bob goes to hospital, but does not know what the medical problem is. She finds the 4-anonymous table containing Bob's record. Bob is 25 years old and lives in suburb with postcode 4352. Bob's record must be record 1, 2, 3 or 4. Based on the table, Alice does not know whether Bob suffers depression or flu. However, she knows Bob did not have flu for a long time. So, Alice knows nearly for sure that Bob suffers depression.

Therefore, *k*-anonymity does not protect individuals from a background knowledge attack.

Machanavajjhala *et al.* presented an *l*-diversity model to enhance the *k*-anonymity model [24]. The *l*-diversity principle is described as the following. A *k*-anonymous group in a disclosed table contains at least *l* well represented values in the sensitive attribute. For example, records 5, 6, 7 and 8 in Table 2 form a 3-diverse group. The records contain three values with the frequencies of 25%, 25% and 50%, and no value is dominant. However, making every *k*-anonymous group include *l* balanced values of a released data set may diminish the usefulness of information in the quasi-identifier in the data set.

A practical model normally trades strong protection with data utility. The *l*-diversity model is strong, but may not be practical. If we try to protect every value in the sensitive attributes, it may be better not to publish either sensitive attributes or the quasi-identifier. Especially for the background knowledge attack, how much knowledge should we assume an adversary has? If an adversary has very strong background knowledge, any published data is not safe.

A more practical approach is not to consider every value in the sensitive attribute as sensitive. For example, people may want to keep depression as private, but not flu and viral infections. If we only have a small number of sensitive vales, a reasonable protection is that the inference confidence from a group of *k*-anonymous records to a sensitive value is below a threshold. This is the basic idea of (*a* ,*k*)-anonymity model [25]. This model keeps the inference confidence to sensitive values lower than *a*, a user defined threshold. This model is simple and effective to prevent some sensitive values from homogeneity attacks. An example of (*a*, *k*)-anonymous table is given in Table 3. In this example, flu and viral infections are not considered as sensitive, the inference confidence from (female & 30-39 & 435*) to depression is 25%.

Another model to prevent homogeneity attack in classification is the template-based model [26]. This model allows users to specify what types of inference channels should be blocked in the released data as templates. The model tries to eliminate the sensitive inferences in a released data set, and to preserve the classification value of the data. The model keeps the confidence of sensitive inferences from a group of individuals to sensitive values lower than the user specified levels. This model is good for users who know exactly what inferences are damaging, but is not suitable for users who do not.

	Quasi identifier			Other attributes	Sensitive attributes
	Gender	Age	Postcode		Diagnosis
1	male	20-29	435*	… … …	flu
2	male	20-29	435*	… … …	flu
3	male	20-29	435*	… … …	flu
4	male	20-29	435*	… … …	flu
5	female	30-39	435*	… … …	depression
6	female	30-39	435*	… … …	flu
7	female	30-39	435*	… … …	viral infections
8	female	30-39	435*	… … …	viral infections

Table 3. A (0.25, 4)-anonymous table

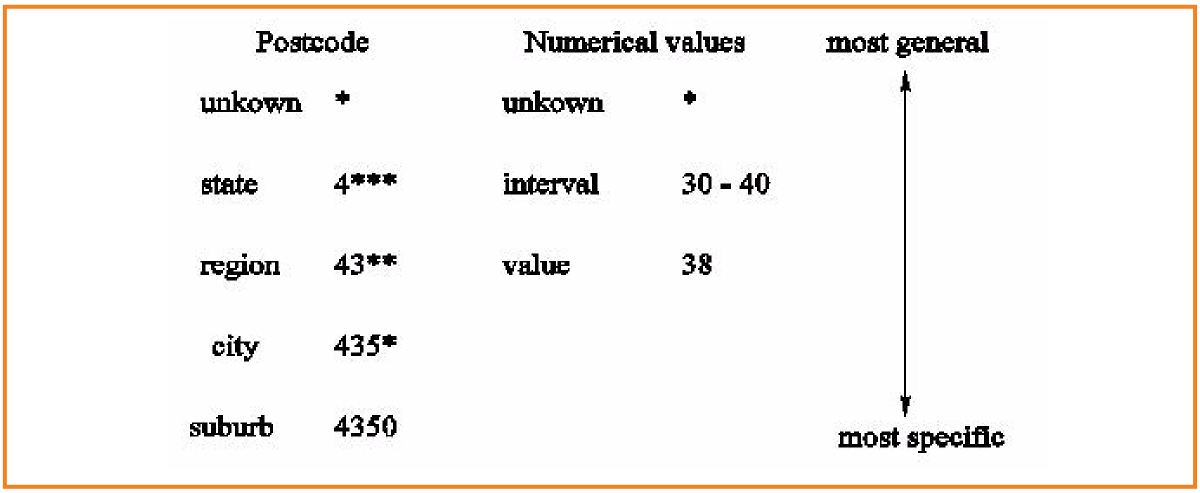

Figure 1: Two examples of domain hierarchies. One for categorical values and one for numerical values.

4. Various *k*-anonymisation methods

There are several different methods to modify a data set to be *k*-anonymous, or so called *k*-anonymisation methods. We outline these techniques in this section.

Generalisation: A common way for *k*-anonymisation is achieved by generalisation — an attribute value is generalised according to its attribute domain hierarchy. For example, date of birth Date/Month/Year is replaced by Month/Year. All attribute domains are in hierarchical structures. Domains with fewer values are more general than domains with more values for an attribute. The most general domain contains only one value. For example, date of birth in Date/Month/Year is a lower level domain, and date of birth in Year is a higher level domain. The most general level of date of birth domain contains value unknown '*' (or Any). Numerical attributes are in a hierarchical structure too. That is {value, interval, *}. Intervals can be determined by users or a machine learning algorithm, say, a discretisation method. As illustrated in Figure 1, 10 year interval level in date birth domain is more general than year level.

Global recoding and local recoding anonymisation methods are two ways to achieve *k*-anonymity. Another name for global recoding is full domain generalisation. In global recoding, the generalisation happens at the attribute domain level. When an attribute value is generalised, every occurrence of the value is replaced by the new generalised value. Many methods are global recoding models, such as [9, 10, 11, 12, 13, 14, 27]. DataFly [9] and Incognito [27] are two typical ones. An advantage of global recoding is that an anonymous view has uniform domains, but it may unnecessarily lose many detailed information.

It is possible to optimise a global recoding method when the quasi-identifier is not large. The optimisation here is in terms of minimising generalisation steps to achieve *k* anonymity. Incognito [27] and *k*-optimise [14] are two examples. However, the time complexity of optimal search is ultimately exponential to the size of quasi-identifier though it is substantially faster than the naive search. These optimal algorithms only perform well on data with small quasi-identifiers.

A major problem with global recoding methods is that they over-generalise the tables and result in many unnecessary distortions. We use an example to show this. Suppose that we have 1000 records, among which only two are uniquely identifiable according to their postcodes. Generalisation of all postcode values to accommodate these two records into a *k*-anonymous table causes too much distortion in the postcode column.

A local-recoding method generalises attribute values at cell level. A generalised attribute value co-exists with the original value. A local recoding method does not over-generalise a table and hence may minimise the distortion of an anonymous view. In the above example regarding over-generalisation problem of the global recoding generalisation, a solution in local recoding will only generalise the two records with other (*k* - 2) records. Optimal local recoding, which is to minimise distortions of a data set for *k*-anonymisation, is NP-hard as discussed in [15, 16]. A local recoding method has to be heuristic, but it generally produces less distortions. More recent work of local recoding k-anonymisation was reported in [25, 28, 29].

In order to make the information loss as low as possible, the software μ-ARGUS [4] aims to reach a good balance between global recoding and local suppression, where attribute values are replaced by a missing value. It starts by recoding some variables globally until the number of unsafe combinations that have to be protected by local suppression is sufficiently low. It allows a user to specify the global recoding interactively, by providing the user with necessary auxiliary information. A user may also decide to let μ-ARGUS eliminate unsafe combinations automatically, which will involve the solution of a complex optimisation problem [4]. μ-ARGUS does not guarantee *k*-anonymity as discovered in [9]. Following a similar strategy as μ -ARGUS for microdata, *t*-ARGUS [4] is developed for the disclosure control of tabular data. *t*-ARGUS works efficiently only on limited number of attributes.

***k*-anonymisation via clustering**: A more general view of *k*-anonymisation is clustering with a constraint of the minimum number of objects in every cluster [30]. A number of methods that deal with numerical attributes approach identity protection by clustering [6, 31] have been proposed. A recent work [32] extends a clustering based method [33] to ordinal attributes, but neither deals with attributes in hierarchical structures. Other work [15, 16] dealing with categorical attributes does not consider attribute hierarchies. Li *et al.* [29] recently presented a method achieving *k*-anonymity in hierarchical attribute structures by local recoding.

5. Comparison of *k*-anonymity model with SMC-based approaches

The definition of privacy-preserving in SMC (Secure Multiparty Computation)-based approaches [21, 34] is different from that of the *k*-anonymity model. It ensures that nothing other than the final data mining results are revealed during the data mining process. In other words, no data miner sees identifications or sensitive attribute values in data. This definition is closer to the definition of security used in SMC techniques initially suggested by Yao [17].

As we mentioned above, both *k*-anonymity and SMC are used in privacy-preserving data mining, but they are quite different in terms of efficiency, accuracy, security and privacy as shown in Figure 2.

The *k*-anonymity provides a formal way of privacy protection by preventing from re-identification of data more than a group of *k* entities, and the data with *k*-anonymity can be used straightway by various parties. However, the application of SMC is inefficient. The notion of computational indistinguishability is crucial to SMC. A distribution of numbers is said to be computationally indistinguishable from another distribution of numbers if no polynomial time program can distinguish the two distributions. Goldreich *et al.* proved that any problem representable as a circuit can be securely solved [35], but the computation cost depends on the input's size. Such a generic method based on circuit evaluation is expensive (even for small inputs) and the computational cost is prohibitive for large inputs. Thus, generic SMC protocols are impractical for the achievement of SMC since large inputs are typically required in data mining. Therefore, the *k*-anonymity model and its variants are more efficient than technniques based on generic SMC protocols. However, in the real world, rather than using a perfect SMC protocol with nothing revealed, more efficient but not completely secure protocols can be used in which we are able to clearly prove what is known and what remains secret. In this direction, there are several research efforts to improve the computation efficiency [21, 34, 36].

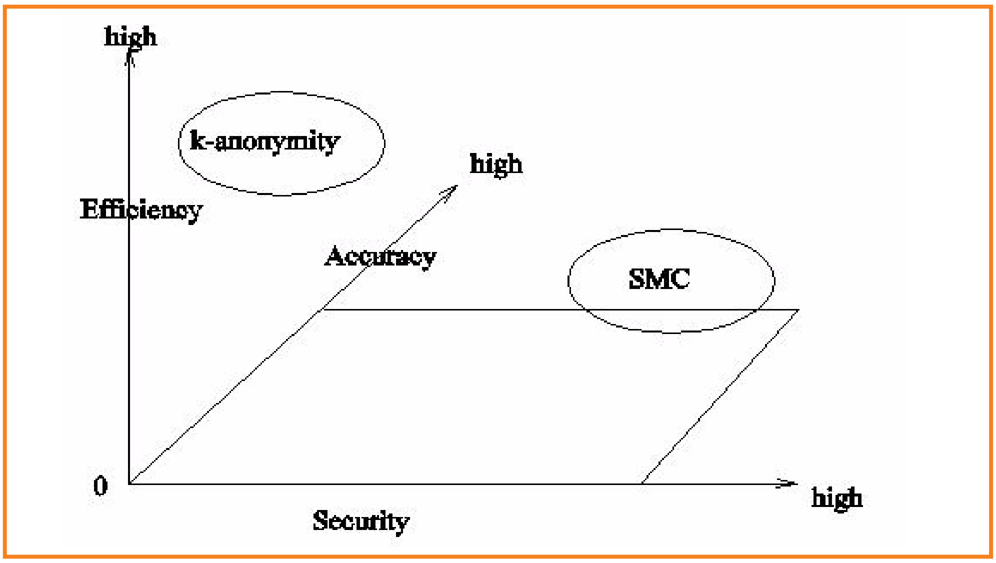

Figure 2: K-anonymity and SMC

A SMC-based solution provides exactly what is revealed. During *k*-anonymisation, the key step is to generalise attribute values. For example, the ages 22 and 25 in Table 1 could be generalised to an interval [20 - 29] in Table 2. Data values in a *k*-anonymous table are modified to satisfy the anonymity feature for privacy protection. This largely trades off the accuracy in data mining. In contrast, a SMC-based solution provides exactly what is revealed.

A general definition of privacy commonly accepted is to protect only the actual data values during transactions. Privacy is preserved if none of the data is known exactly. A stronger definition of privacy has initiated by the cryptographic community. For example, Yao [18] proposed a semantic definition of it: *information communicated between parties won't enable one party to compute any polynomial time predicates with greater accuracy than they could without the communication.* On the other hand, the *k*-anonymity model aims to protect just the (exact) actual data values. Privacy is preserved in the *k*-anonymity model when no exact values are learned. The security definition of the *k*-anonymity model is much weaker than the one in the SMC model.

Almost all the SMC-based approaches may suffer privacy inference vulnerability [37], because they do not consider the privacy risk caused by the release of final result [38]. For example, via a SMC-based approach, the following association rule from a binary table is released.

$a_1 \wedge a_2 \rightarrow a_3$ (support = 99; confidence = 99%]

It means that 99 individuals have a_1 and a_2 and a_3 simultaneously. According to the association rule definition [38], one may easily derive that 100 (support confidence = 99 / 99.0%) individuals have both a_1 and a_2. This means that one and only one individual in the data has a_1 and a_2 but not a_3. Thus $a_1 \wedge a_2 \wedge \neg a_3$ is unique. This association rule could be used a key for a linking attack. As a comparison, enhanced *k*-anonymity models focus on eliminating this kind of privacy threats.

6. Conclusions and discussions

In this paper, we have introduced *k*-anonymity model and discussed how it protects private information in data. We have outlined typical problems associating with the *k*-anonymity model and considered some enhanced models to overcome the problems. We have summarised major techniques implementing *k*-anonymisation. We also have simply compared the *k*-anonymity model with the Secure Multi-party Computation (SMC)-based techniques, in terms of efficiency, accuracy and security.

There are several problems of the *k*-anonymity model and its variants.

- One is how to determine quasi-identifiers during *k*-anonymisation of a data set. In other words, which kinds of other data sources would be used to construct a linking attack? The problem is arguably the same as which kind of background knowledge an attacker would have. This problem may be difficult for both data custodians and privacy-preserving technique developers to answer.

- The second one is how to choose an optimal parameter *k* for the *k*-anonymity model for a given data set. A similar problem is what the best trade-off between privacy protection and data accuracy is.

- Do these enhanced *k*-anonymity models, such as *l*-diversity or (alpha, *k*)-anonymity, guarantee privacy? If not, is the underlying principle of SMC able to be used to further enhance these models? Is there a model which guarantees privacy theoretically?

- As for the potential privacy inference vulnerability of the SMC-based privacy preserving techniques discussed in Section 6, can the *k*-anonymity model or its variants be used to enhance these SMC-based techniques?

- A data set can be anonymised in different ways. How do we measure the quality of various k-anonymous tables?
- When more than one k-anonymous table are released due to the update, are there any risks for privacy breaching?

Most of these problems are crucial to the real world applications of *k*-anonymous data releasing. We leave them as future research directions.

Acknowledgements

This work has been supported by ARC discovery project DP0774450.

References

1. Adam NR and Wortmann JC, Security-control methods for statistical databases: A comparative study. *ACM Comput. Surv.*, 21(4):515–556, 1989.

2. Cox L, Suppression methodology and statistical disclosure control. *J. American Statistical Association*, 75:377–385, 1980.

3. Willenborg L. and Waal T, Statistical disclosure control in practice. *Lecture Notes in Statistics*, 111, 1996.

4. Hundepool A and Willenborg L, μ-and *t*-ARGUS: software for statistical disclosure control. In *Third international seminar on statsitcal confidentiality*, pages 142–149, Bled, 1996.

5. Agrawal R and Srikant R, Privacy-preserving data mining. In *Proc. of the ACM SIGMOD Conference on Management of Data*, pages 439–450. ACM Press, May 2000.

6. Agrawal D and Aggarwal CC, On the design and quantification of privacy preserving data mining algorithms. In *PODS '01: Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 247–255, New York, NY, USA, 2001. ACM Press.

7. Rizvi S and Haritsa J, Maintaining data privacy in association rule mining. In *Proceedings of the 28th Conference on Very Large Data Base (VLDB02)*, pages 682–693. VLDB Endowment, 2002.

8. Samarati P and Sweeney L, Generalizing data to provide anonymity when disclosing information (abstract). In *Proceedings of the Seventeenth ACM SIGACTSIGMOD- SIGART Symposium on Principles of Database Systems*, page 188, 1998.

9. Sweeney L, Achieving k-anonymity privacy protection using generalization and suppression. *International journal on uncertainty, Fuzziness and knowledge based systems*, 10(5):571 – 588, 2002.

10. Samarati P, Protecting respondents' identities in Microdata release. *IEEE Transactions on Knowledge and Data Engineering*, 13(6):1010–1027, 2001.

11. Iyengar VS, Transforming data to satisfy privacy constraints. In *KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 279–288, 2002.

12. Wang K, Yu PS, and Chakraborty S, Bottom-up generalization: A data mining solution to privacy protection. In *ICDM04: The fourth IEEE International Conference on Data Mining*, pages 249–256, 2004.

13. Fung BCM., Wang K, and Yu PS, Top-down specialization for information and privacy preservation. In *ICDE05: The 21st International Conference on Data Engineering*, pages 205–216, 2005.

14. Bayardo R and Agrawal R, Data privacy through optimal k-anonymization. In *ICDE05: The 21st International Conference on Data Engineering*, pages 217–228, 2005.

15. Aggarwal G, Feder T, Kenthapadi K, Zhu A, Panigrahy R, and Thomas D, Achieving anonymity via clustering in a metric space. In *Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems (PODS 06)*, pages 153–162, 2006.

16. Meyerson A and Williams R, On the complexity of optimal k-anonymity. In *PODS04: Proceedings of the twenty fourth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 223–228, 2004.

17. Yao A, Protocols for secure computations. In *Proceedings of Twenty-third IEEE Symposium on Foundations of Computer Science*, pages 160–164, Chicago, Illinois, 1982.

18. Yao A, How to generate and exchange secrets. In *Proceedings of Twenty-seventh IEEE Symposium on Foundations of Computer Science*, pages 162–167, Toronto, Canada, 1986.

19. Lindell Y and Pinkas B, Privacy preserving data mining. *Journal of Cryptology*, 15(3):177–206, 2002.

20. Vaidya J and Clifton C, Privacy-preserving data mining: why, how, and when. *Security & Privacy Magazine, IEEE*, 2(6):19 – 27, Nov.-Dec. 2004.

21. Wright R and Yang Z, Privacy-preserving Bayesian network structure computation on distributed heterogeneous data. In *KDD'04: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 713–718, New York, NY, USA, 2004. ACM Press.

22. Kantarcioglu M, Jin J, and Clifton C,When do data mining results violate privacy? In *KDD '04: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 599–604, New York, NY, USA, 2004. ACM Press.

23. Sweeney L, k-anonymity: a model for protecting privacy. *International journal on uncertainty, Fuzziness and knowledge based systems*, 10(5):557 – 570, 2002.

24. Machanavajjhala A, Gehrke J, and Kifer D, *l*-diversity: privacy beyond *k*-anonymity. In *Proceedings of the 22nd International Conference on Data Engineering (ICDE06)*, page 24, 2006.

25. Wong R, Li J, Fu A, and Wang K, (alpha, k)-anonymity: An enhanced kanonymity model for privacy-preserving data publishing. In *Proceedings of the twelfth ACM SIGKDD international conference on knowledge discovery and data mining (KDD)*, pages 754–759, 2006.

26. Wang K, Fung BCM, and Yu PS, Template-based privacy preservation in classification problems. In *ICDM05: The fifth IEEE International Conference on Data Mining (ICDM'05)*, pages 466 – 473, 2005.

27. LeFevre K, DeWitt DJ, and Ramakrishnan R, Incognito: Efficient full-domain k-anonymity. In *SIGMOD '05: Proceedings of the 2005 ACM SIGMOD international conference on Management of data*, pages 49–60, 2005.

28. LeFevre K, DeWitt D, and Ramakrishnan R. Mondrian multidimensional k-anonymity. In *Proceedings of the 22nd International Conference on Data Engineering (ICDE 2006)*, page 25, 2006.

29. Li J, Wong R, Fu A, and Pei J, Achieving k-anonymity by clustering in attribute hierarchical structures. In *Proceeding of 8th International Conference on Data Warehousing and Knowledge Discovery*, pages 405–416, 2006.

30. Aggarwal G, Feder T, Kenthapadi K, Motwani R, Panigrahy R, Thomas D, and Zhu A.. Anonymizing tables. In *ICDT05: Tenth International Conference on Database Theory*, pages 246–258, 2005.

31. Aggarwal CC, On k-anonymity and the curse of dimensionality. In *VLDB '05: Proceedings of the 31st international conference on Very large data bases*, pages 901–909. VLDB Endowment, 2005.

32. Domingo-Ferrer J and Torra V, Ordinal, continuous and heterogeneous kanonymity through microaggregation. *Data Mining and Knowledge Discovery*, 11(2):195–212, 2005.

33. Domingo-Ferrer J and Mateo-Sanz JM, Practical data-oriented microaggregation for statistical disclosure control. *IEEE Transactions on Knowledge and Data Engineering*, 14(1):189–201, 2002.

34. Vaidya J and Clifton C, Privacy-preserving k-means clustering over vertically partitioned data. In *The Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 24–27, Washington, DC, 2003. ACM Press.

35. Goldreich O, Micali S, and Wigderson A, How to play any mental game. In *STOC '87: Proceedings of the nineteenth annual ACM conference on Theory of computing*, pages 218–229, New York, NY, USA, 1987. ACM Press.

36. Gilburd B, Schuster A, and Wolff R, k-TTP: a new privacy model for large scale distributed environments. In *Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 563–568. ACM Press, 2004.

37. Farkas C and Jajodia S, The inference problem: a survey. *SIGKDD Explor. Newsl.*, 4(2):6–11, 2002.

38. Atzori M, Bonchi F, Giannotti F, and Pedreschi D, k-anonymous patterns. In *Proceedings of the 9th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD05)*, pages 10–21, 2005.

Correspondence

Dr. Jiuyong Li
School of Computer and Information Science,
University of South Australia, Mawson Lakes
South Australia, Australia, 5095
Email: jiuyong.li@unisa.edu.au
Tel: +61 8 830253898
Fax: +61 8 83023381